图书情报工作
第 62 卷 第 10 期 2018 年 5 月


# 基于非负矩阵分解的技术主题演化分析

王园园 赵亚娟
中国科学院文献情报中心 北京 100190 中国科学院大学经济管理学院 北京 100190

**摘要**：[目的/意义]分析技术主题演化过程可以梳理技术发展脉络，对于发展创新、预测技术发展趋势具有重要意义，但是从语义角度分析技术主题演化轨迹的研究较少。因此，从语义的角度出发，分析技术主题演化过程。[方法/过程]提出基于非负矩阵分解的改进的动态非负矩阵分解模型对专利文本进行动态主题建模，并利用 TextRank 算法抽取名词短语进行标注，增强所抽取技术主题的可解释性。在此基础上，利用词向量的方式计算技术演化轨迹，并进行可视化展示。[结果/结论]对 2002 年、2005 年、2008 年、2011 年和 2014 年的五方专利进行实证分析，识别出 65 个技术主题及其演化轨迹，表明方法的可行性。

**关键词**：技术主题演化 非负矩阵分解 主题模型 动态主题分析
**分类号**：G254.11
**DOI**：10.13266/j.issn.0252-3116.2018.10.013



现代社会技术的发展日新月异，产业间的技术流动、技术合作以及不同产业间的技术交融不断增强，技术关联愈发紧密，一个产业的技术进步与其它产业的技术变化息息相关[1-2]。作为社会创新的主体，企业必须面临通过不断地创新以持续研发新产品的挑战。因此，技术的复杂性与多样性与日俱增，技术创新的步伐加快，强度提升，技术发展过程中的不确定性也在不断增强[3]。技术主题分析是专利情报分析的重要内容，主要分为技术主题分布和技术主题演化分析两个方面，其中技术主题分布侧重于技术主题的静态特征，技术主题演化分析内涵较为丰富，包含技术主题演变过程分析、技术发展趋势预测和新兴技术主题发现等内容[4]。理解技术主题演化机制对于发展创新具有重要意义。

专利作为技术、法律以及商业信息的载体，是技术主题演化研究的重要数据资源。日益增长的专利数量和日益复杂的技术给技术主题演变分析带来了巨大的挑战。随着文本挖掘、语义分析技术的发展，主题模型（如 LDA、NMF 等）在众多领域（如社交媒体、科学文献等）得到了广泛、成熟的应用，极大地提高了人们挖掘、理解非结构化文本数据语义信息的效率，这也为技术主题演变分析研究提供了一种有价值的思路，即从专利的文本内容分析技术主题演变的动态过程。本文从非结构化文本数据的角度，基于非负矩阵分解（NMF）提出改进的动态非负矩阵分解的方法，将专利文本划分为不同的时间窗口并抽取窗口主题，再基于窗口主题抽取动态主题，以探究技术主题的动态演变过程。

## 1 相关文献综述

专利是技术主题演化研究的主要数据来源，本文根据现有研究对专利信息的利用方式的不同，将技术主题演变研究分为三类：基于专利分类的分析方法、基于专利引文的分析方法和基于专利文本内容的分析方法。

专利分类是根据专利揭示的技术内容所提供的一种简易和通用的技术分类系统[5]。基于专利分类号的技术主题分析主要有统计分析和共分类分析。专利共分类是指不同的专利分类号码（如 IPC）在同一件专利中共现，表明不同的技术方向之间存在一定的联系，可以基于这种联系分析技术主题。K. Suzuki 等[6]采用专利 IPC 分类号共现的方法来研究技术发展中的融合。S. Jeong 等[7]通过 Jaccard 系数来研究 IPC 共现关系的强弱，并分析了共现网络的密度等特征，以分析不同技术主题随时间的变化及技术融合的主要类型。W. S. Lee 等[8]对 IPC 的共现网络进行链路预测（link prediction）分析以预测未来可能产生的新兴技术主题，

作者简介：王园园（ORCID：0000-0003-1079-0766），硕士研究生；赵亚娟（ORCID：0000-0003-3501-8131），研究员，博士，硕士生导师，通讯作者，E-mail：zhaoyj@mail.las.ac.cn。
收稿日期：2017-10-30 修回日期：2018-01-22 本文起止页码：94-105 本文责任编辑：王善军

并利用主题分析抽取关键词来识别未来可能的新兴领域。B. Huang 等[9]利用关联规则分析的方法对信息技术和生物技术两大技术领域的 IPC 共现进行了分析,从支持度(support)、置信度(confidence)和提升度(lift)三个方面分析了技术主题的特征。

基于专利引文的分析方法对不同专利文献之间以及专利文献与科学文献之间的引用关系进行分析[10]。引用关系反映了技术的流动,通过构建引用关系网络,可以分析技术主题演变的轨迹。P. L. Chang 等人[11]以专利引用关系为基础,结合层次聚类和非层次聚类方法,将目标领域专利聚成三个类簇,为每个类簇生成技术主题,并构建了每个类簇内部技术之间关系的网络图。C. Choi 等人[12]提出一种基于专利引用网络的技术分析方法,可以识别技术主题演化路径。Y. GEUM 等[13]、翟东升等[14]则根据技术类别之间引用网络的知识流来分析技术融合。除基于直接引用关系的专利引文网络外,还有以共被引关系和引文耦合关系作为专利的技术主题相似度构建的专利网络[15-16]。总体来说,当前基于专利引文的技术主题分析方法主要有三类[17]:基于专利引文关系进行聚类操作,并进一步分析技术主题的演化情况[12,15-16];通过识别专利引文网络中的知识流动主路径以绘制技术主题演化轨迹[18-19];采用社会网络分析的方法来评价技术主题演化阶段[20-21]。

基于专利分类和专利引文的技术主题演化分析方法虽然能从宏观角度发现技术发展趋势,但无法展现技术主题的具体演变细节。基于专利文本的分析弥补了这一不足,挖掘专利文本中隐藏的信息逐渐成为技术主题演化分析的主要手段之一[22]。当前基于专利文本的技术主题分析主要有词频分析法和关键词共现分析等。栾春娟[23]基于主题词共现分析方法和社会网络分析方法,绘制太阳能技术领域共现网络的演进过程。韩红旗等[24]提出专利技术特征词共现的战略图分析方法研究技术主题的演化情况。S. H. Chen 等[25]对不同时间窗口的专利文本进行聚类,结合专利引用网络分析技术主题的演化过程。

本文从语义的角度出发,提出基于 NMF 的改进模型,以实现技术主题演变分析。NMF 是一种将非负矩阵分解、降维为非负因子的无监督方法,被广泛应用于图像处理[26]、文本语料潜在主题的抽取[27]等领域。将 NMF 直接用于主题建模是静态的,无法反映所抽取的主题在时间尺度上的演变情况。本文基于 NMF 提出一种改进的动态非负矩阵分解(Dynamic NMF)的方法,首先将专利文本划分为不同的时间窗口,分别对每个时间窗口的专利文本利用 NMF 进行主题建模,得到窗口主题模型,然后基于窗口主题模型再次利用 NMF 进行主题建模,得到动态主题模型,最终以反映动态主题在不同时间窗口的演变情况,从语义的角度揭示技术主题的动态演变特征。

## 2 研究方法

本文提出的总体研究框架如图 1 所示,主要分为四个步骤:①对抽取的专利文本数据训练词向量,得到的词向量用于后续的主题一致性评价和主题相似度的计算;②对抽取的专利文本数据利用动态 NMF 进行动态主题建模,得到动态主题和窗口主题,其中主题个数的确定是利用基于词向量的主题一致性评价指标;③通过 TextRank 算法抽取的名词短语对抽取的主题进行短语标注,增强主题的可解释性;④计算主题相似度,识别技术主题演变轨迹。下文将对框架中的重要步骤进行介绍。

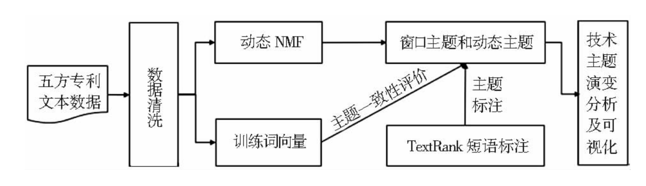



图 1 研究总体框架

### 2.1 非负矩阵分解(NMF)

给定包含 $n$ 篇文档的语料库,首先构建文档 - 词矩阵 $A \in R^{n \times m}$,其中,$m$ 表示语料库词表的长度。对矩阵 $A$ 进行 NMF,结果产生 $k$ 秩(rank)的降维近似,这种近似是两个非负因子乘积的形式,即 $A \approx WH$。NMF 的目标是最小化 $A$ 与 $WH$ 之间的重构误差(reconstruction error)。因子 $H \in R^{k \times m}$ 的行可以解释为 k 个主题(topic),每个主题定义为词表中 m 个词的非负权重。对每一行按照词的权重进行排序,即可得到每个主题的 *top n* 词表示。矩阵 $W \in R^{n \times k}$ 的列表示 n 篇文档对于每个主题的权重,基于此可以将文档与对应的主题联系起来。NMF 算法通常以随机因子进行初始化,导致算法收敛于不同的局部最优,进而导致算法结果的不稳定性。本文使用非负双重奇异值分解(NNDSVD)[28]生成初始化因子,以提升所抽取主题的质量。

此外,主题模型中一个关键的参数就是主题个数 $k$,直接决定了主题抽取的结果。$k$ 值过小,则抽取的主题过于宽泛;$k$ 值过大,将导致过多的、高相似度的主题。J. Chang 等[29]将主题一致性(topic coherence)用

于不同 $k$ 值主题之间的比较,主题一致性假设同一主题中的词是相关的,并通过这种相关性来评价主题的质量。D. O'Callaghan 等[30]提出基于 Word2Vec 的主题一致性评价方法(Topic Coherence via Word2Vec,TC-W2V),该方法通过评价主题 *top n* 词的相关性来评价总体主题的一致性。具体来讲,TC-W2V 通过 Word2Vec[31]来计算词表中词的向量表示,然后通过余弦相似度计算同一主题中词对之间的相关性。通常,主题中词之间的相似度越高,则主题的语义一致性就越高。本文利用 TC-W2V 确定 k 值,如式(1)和(2)所示,每个主题由前 $t$ 个词表示,对于单个主题 $t_h$,一致性为前 $t$ 个词两两之间余弦相似度的均值,其中词的向量表示由 Word2Vec 计算获得:

$$coh(t_h) = \frac{1}{\binom{t}{2}} \sum_{j=2}^{t} \sum_{i=1}^{j-1} cos(wv_i, wv_j) \quad (1)$$

对于由 $k$ 个主题构成的主题模型 $T$,总体的一致性得分由每个主题一致性的均值给出:

$$\mathrm{coh}(T) = \frac{1}{k} \sum_{h=1}^{k} \mathrm{coh}(t_h) \quad (2)$$

给定区间$[k_{min}, k_{max}]$,可以通过 TC-W2V 的最大值确定最优的 $k$ 值。

### 2.2 动态非负矩阵分解

动态非负矩阵分解是一个两层的非负矩阵分解,具体来讲,对于具有时序特征的专利文本(比如以年为单位),本文首先利用非负矩阵分解对每个固定时间窗口的专利文本进行主题建模,然后将抽取的主题视为文本,再次利用非负矩阵分解对每个时间窗口的输出结果进行主题建模,以抽取所有时间窗口的专利文本中所蕴含的动态技术主题。动态非负矩阵分解的过程如图 2 所示:

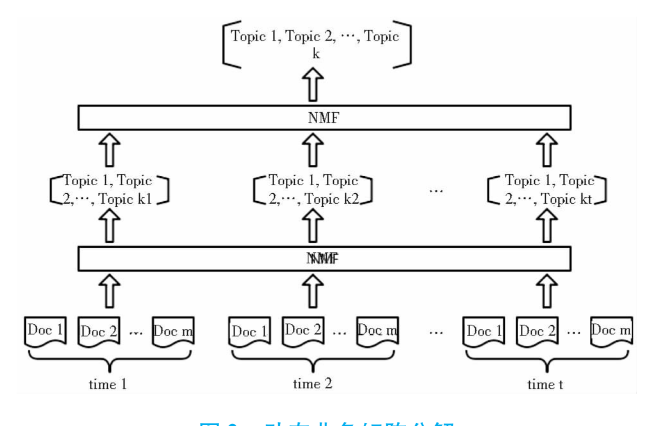



图 2　动态非负矩阵分解

第一层面对时序数据,首先要将数据划分到固定长度的时间窗口中。关于时间窗口的划分,主要有重叠时间窗口划分[32]和非重叠时间窗口划分[33]两种形式。重叠时间窗口划分的方式容易忽略一些存在周期较短的主题,也忽略了主题在每个时间点的状态。本文采用非重叠的方式将专利文档划分到 $\tau$ 个时间窗口 $\{T_1, \cdots, T_\tau\}$ 中,对每一个时间窗口 $T_i$ 应用 NMF,产生包含 $k_i$ 个窗口主题(window topic)的窗口主题模型 $M_i$,其中参数 $k_i$ 由式(2)确定。第一层产生了连续的窗口主题模型$\{M_1, \cdots, M_\tau\}$。

第二层对于每个窗口主题模型中的因子 $H_i$,将 $H_i$ 的行(即 $k_i$ 个窗口主题)看成是主题文档(主题是用词表示的,因此也可看作是文档),即可构建原始语料的压缩表示。主题–词矩阵 $B$ 的构建方式如下:①构建空矩阵 $B$;②对每个窗口主题模型 $M_i$:对 $M_i$ 中的每个窗口主题,从对应的 NMF 因子 $H_i$ 的行向量中选择前 $t$ 个词,将其它词的权重置为 0,将此向量作为新行添加到 $B$ 中;③所有窗口主题模型中的主题添加完成之后,删除 $B$ 中全为 0 的列(未曾在任何窗口主题前 $t$ 个词中出现过的词)。

矩阵 $B$ 的大小为 $n' \times m'$,其中 $n' = \sum_{i=1}^{\tau} k_i$ 是主题文档的数量,$m' \ll m$ 是上述步骤 3 余下的词的子集。保留主题文档中的前 $t$ 个词实际上利用了每个时间窗口中有代表性的词,并且排除了每个窗口主题中的低意义词,最终降低第二次因子分解过程的计算代价。对矩阵 $B$ 进行第二层 NMF 抽取 $k'$个动态主题(dynamic topic),每个动态主题都会与多个时间窗口联系起来。矩阵 $B$ 的分解过程与矩阵 $A$ 的分解过程一样,TW-W2V 一致性度量用来确定参数 $k'$的值。分解的结果 $B \approx UV$ 可以解释为:因子 V 每一行的 *top n* 词用来表示动态主题;因子 $U$ 的列值表示每个窗口主题与每个动态主题的相关程度。

将动态主题与窗口主题联系起来能够追踪主题随时间的演变。首先,基于因子 $U$ 每一行的值,将每一个窗口主题与其所属权重最大的动态主题进行关联。同理,可将专利文档与窗口主题进行关联,进而将动态主题与专利文档关联起来。两层 NMF 主题模型过程的输出结果包含:①$\tau$ 个窗口主题模型,每个窗口主题模型包含 $k_i$ 个窗口主题,每个窗口主题有一组与之关联的文档,且用前 $t$ 个词表示;②$k'$个动态主题,每个动态主题与一组窗口主题关联,并且有一组与之关联的文档。

## 3　数据

当前技术主题演化研究多集中于某个技术领

域[14,34－35]，缺乏对全领域的研究。五方专利是指同时在中美欧日韩五国申请授权的专利，一般专利质量较高，覆盖的技术领域较为广泛，具有代表性。本文使用德温特专利数据库作为专利信息的数据源，检索2002、2005、2008、2011 及 2014 年五个年份的五方专利数据，五年数据分别为 16 500 条、25 221 条、25 866 条、24 184 条、20 947 条，共计 112 718 条记录，如图 3 所示。检索时间为2016 年 10 月 28 日。

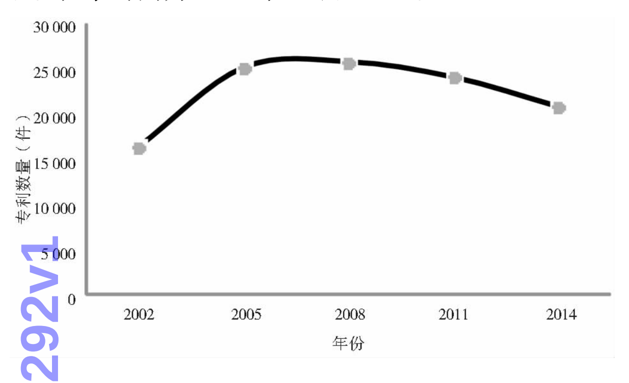



图 3　专利数量统计

本文按照年份将数据集划分为五个时间窗口，对每个时间窗口 $T_i$，按如下方式构建文档－词矩阵 $A_i$：①抽取 $T_i$ 窗口所有记录的标题和摘要信息作为专利文档；②对文档进行词令化(tokenization)，对词进行大小写转换，并进行词形还原(lemmatization)；③去除词长<3 的词令及停用词(stop words)，其中，停用词表的构建基于德温特专利记录的特点，添加了如“advantage”、“description”等词，共计 831 个停用词；④低频词往往不具备足够的代表性，而高频词所表征的意义又过于宽泛，因此为平衡所抽取主题的区分度和涵盖的范围，本文去除了词频小于 20 的低频词以及在超过 60% 的文档中出现的高频词；⑤构建文档－词矩阵 A_i，并计算 TF-IDF 权重。

处理结果显示五个时间窗口分别包含 5 107、6 414、7 028、6 627、6 481 个词项。

## 4　结果与分析

### 4.1　确定主题数目 *k*

如上文所述，本文利用主题一致性指标 TC-W2V 自动确定 *k* 值。首先利用全部的专利文本数据训练 Word2Vec，每个主题的前 20 个词用来计算 TC-W2V 值。在每个时间窗口，在兼顾程序运行效率的情况下，设定主题个数的取值范围 $k \in [80,180]$，步长为 5，生成不同主题个数的窗口主题模型，确定最优的 *k* 值，即主题一致性 TC-W2V 最高的 *k* 值。图 4 展示了 2014 年不同 *k* 值窗口主题模型的 TC-W2V 得分，$k = 110$ 时 TC-W2V 取得最大值。同理，2002、2005、2008、2011 四个时间窗口的最优 *k* 值分别为 105、120、100、110。动态主题模型主题个数的取值范围设定为 $k \in [50,150]$，步长为 5，最优 *k* 值为 65。本文共抽取动态主题 65 个，分别用 *D*01，*D*02，…，*D*65 表示，窗口主题 545 个，分别用“年份＋主题序号”表示，如 2002_01，2002_02，…，2014_110。

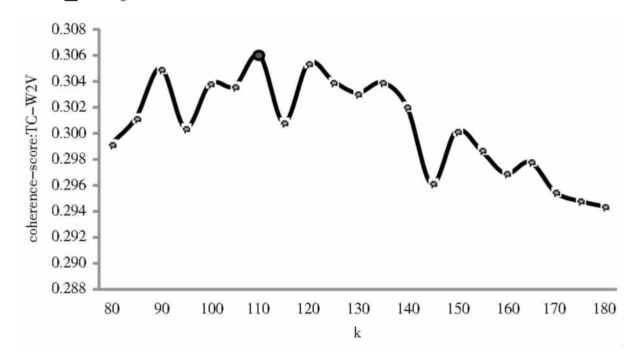



图 4　2014 年主题一致性 TC-W2V

### 4.2　技术主题演化分析

本文通过提出的动态非负矩阵分解模型抽取出 65 个动态主题，为进一步分析这些技术主题的演化过程，利用主题强度和技术融合度两个指标作为筛选标准。其中，主题强度是指与该主题关联的专利的数量，技术融合度量方式有多种，不同度量方式从不同角度揭示了技术融合的不同特征[36]。熵通过度量一个技术方向在不同技术类别上的分布情况来度量技术融合度，如 E. J. Han 和 S. Y. Sohn[37] 以及 Y. Cho 和 M. Kim[3] 等将专利分类信息(IPC)的熵(entropy)用来度量某一技术主题的技术融合度，本文也沿用这一做法。具体来说，本文将识别的技术主题与专利联系起来(将每件专利与其对应权重最大的主题进行关联)，进而将识别的技术主题与专利所包含的分类信息(IPC)关联起来，得到与该技术主题关联的 IPC 的频次分布信息。基于该技术主题的 IPC 分布信息，即可计算熵。本文通过统计 4 位 IPC 来度量技术融合度，如式(3)所示，其中 $P(x_i)$ 表示某一技术类别(4 位 IPC)出现的频率：

$$H(X) = -\sum_i P(x_i)\log(P(x_i)) \qquad (3)$$

本文所识别的 65 个动态主题的主题强度和技术融合度如表 1 所示，为进一步分析所识别技术主题的演化过程，以主题强度和技术融合度都较高的动态主题 D64 和 D59 为例，分析技术主题演化过程。

表 1 动态主题的主题强度和技术融合度(按技术融合度降序)

| 动态主题 | 技术融合度 | 主题强度 | 动态主题 | 技术融合度 | 主题强度 | 动态主题 | 技术融合度 | 主题强度 |
|---|---|---|---|---|---|---|---|---|
| D59 | 5.38 | 8 542 | D12 | 4.52 | 441 | D20 | 3.59 | 761 |
| D64 | 5.34 | 5 337 | D44 | 4.52 | 852 | D38 | 3.58 | 607 |
| D54 | 5.33 | 4 822 | D28 | 4.48 | 604 | D07 | 3.53 | 5 884 |
| D24 | 5.27 | 959 | D52 | 4.47 | 738 | D45 | 3.53 | 1 089 |
| D43 | 5.27 | 758 | D10 | 4.42 | 2 633 | D36 | 3.46 | 574 |
| D19 | 5.19 | 632 | D25 | 4.41 | 3 511 | D30 | 3.46 | 1 630 |
| D23 | 5.16 | 1 943 | D14 | 4.40 | 1 007 | D15 | 3.39 | 94 |
| D18 | 5.16 | 893 | D27 | 4.33 | 731 | D55 | 3.34 | 3 887 |
| D33 | 5.04 | 654 | D48 | 4.29 | 613 | D35 | 3.34 | 437 |
| D47 | 4.95 | 3 789 | D16 | 4.23 | 954 | D60 | 3.29 | 4 041 |
| D31 | 4.90 | 739 | D53 | 4.12 | 664 | D34 | 3.27 | 610 |
| D63 | 4.84 | 2 567 | D06 | 4.08 | 1 434 | D29 | 3.25 | 1 771 |
| D21 | 4.82 | 528 | D46 | 4.06 | 927 | D13 | 3.25 | 708 |
| D58 | 4.79 | 1 291 | D26 | 4.05 | 1 059 | D49 | 3.22 | 1 265 |
| D61 | 4.74 | 2 455 | D22 | 4.05 | 644 | D50 | 3.11 | 1 077 |
| D57 | 4.72 | 1 443 | D42 | 4.04 | 1 885 | D11 | 3.04 | 2 716 |
| D32 | 4.71 | 2 408 | D39 | 3.98 | 425 | D04 | 3.01 | 6 979 |
| D01 | 4.69 | 710 | D41 | 3.94 | 1 528 | D09 | 2.82 | 609 |
| D65 | 4.64 | 1 850 | D56 | 3.94 | 2 278 | D03 | 2.72 | 760 |
| D37 | 4.56 | 940 | D08 | 3.90 | 1 916 | D51 | 2.56 | 682 |
| D40 | 4.53 | 740 | D62 | 3.76 | 1 133 | D17 | 2.52 | 893 |
| D02 | 4.53 | 3 919 | D05 | 3.72 | 1 748 | | | |



主题模型生成的主题通常用与主题最相关的 top n 词表示[27,38-39]。然而词的意义较为宽泛,与之相比,短语的语义表达更加完整精确,尤其对于专利文档中的技术术语而言。主题标注(topic labeling)则有助于人们理解主题的含义[40-43]。本文利用 TextRank[44] 算法对 NMF 生成的主题进行短语标注。TextRank 计算每个词的重要性,如果文档中两个相邻的词都是重要的则构成短语。再结合词性标注进行句法过滤,即得名词短语。

动态主题 D64 和 D59 如表 2 所示,其中 NMF 对应的主题内容表示由 NMF 主题模型生成的与该主题最相关的前 20 个词,TextRank 对应的主题内容表示由 TextRank 算法抽取的名词短语所标注的内容。通过两者的对比可以看出,虽然 NMF 直接生成的主题词已经能够较为清晰地表达该主题的内容,但基于单词的表达仍显得较为宽泛,使得主题蕴含的语义不够完整、准确。名词短语则使得主题所蕴含的语义表达的更为清晰、准确和完整。为进一步挖掘所抽取技术主题表达的内容,笔者抽取与该主题最相关的几篇专利,用来辅助主题语义内涵的理解。容易看出,动态主题 D64 所表达的内涵是汽车生产与制造相关的技术,动态主题 D59 所表达的是电气设备相关技术。

表 2 动态主题 D64 和 D59

| 动态主题 | | 主题内容 | 释义 |
|---|---|---|---|
| D59 | NMF | section; housing; connector; contact; assembly; end; electrical; body; connection; cable; conductive; module; member; tube; plug; wall; conductor; connecting; board; structure | 电气设备相关技术 |
| | TextRank | four-line wire structure; layer comprises polyamide; evaporating device; high-friction section; erected portion; process fluid forms; use nozzle; includes link; olap cube determination module; provide directional; ground material; separate large-scale surface; body performance; tomogram selection unit; end flaps; nuclear industry; superconductor layers; cross-sectional microscopic structure; structure; comprises | |
| D64 | NMF | magnetic; vehicle; motor; shaft; drive; coil; rotor; magnet; wheel; electric; gear; core; field; stator; bearing; speed; permanent; rotation; engine; machine | 汽车生产与制造相关技术 |
| | TextRank | vehicle vision; magnet unit; magnetic radial bearings surround; cathode layer comprises; outer surface portion; vertical seal device; motor car; use hybrid operating; drive connector; includes variable-pitch; monomeric forms; controls displacement range; rotatory shaft; provides makeup; second vehicle; vehicle; aramid fiber material; hydrodynamic elements; unit | |

4.2.1　技术主题演化定量分析　动态主题 D59“电气设备相关技术”和 D64“汽车生产与制造相关技术”的专利数量随时间的演化如图 5 所示，专利数量越多，表明该技术越热门、越重要。从图 5 可以看出，在 2002 年到 2011 年之间，“汽车生产与制造相关技术”的专利数量不断增加，表明该技术在这一时间段内得到了快速的发展，而在 2014 年，该技术领域相关的专利数量出现急剧下滑的趋势，表明该技术领域由快速发展阶段进入平稳发展的阶段。根据图 3 所有五方专利数量的统计，在 2014 年总体五方专利数量也有下滑的趋势，所以从另一个角度来看，汽车动力与制造相关技术专利数量的下滑也有可能是受到总体五方专利数量下滑的影响。在 2002 年到 2005 年之间，“电气设备相关技术”专利数量增长迅速，之后趋于平稳状态，表明技术发展放缓，进入平稳发展阶段。

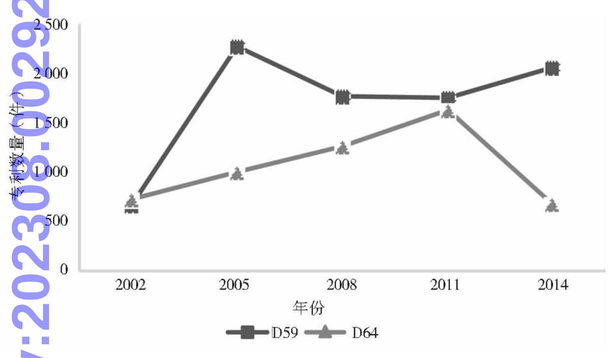



图 5　技术主题专利数量随时间的演化

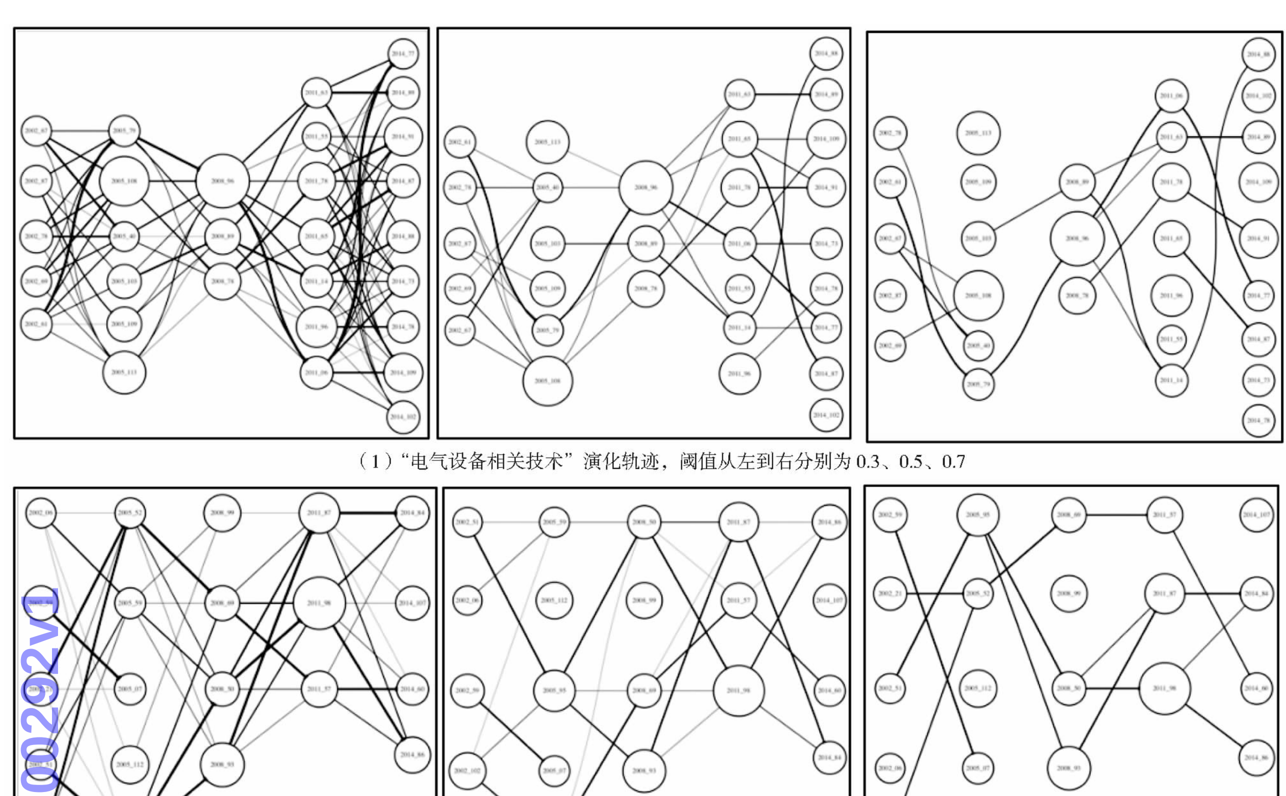
4.2.2　技术主题演化内容分析　基于 NMF 的动态主题分析技术不仅能够揭示文档集蕴含的内容，而且能够揭示动态主题在不同时间窗口的演变情况。本文利用 Word2Vec 计算相邻时间窗口主题之间的相似度，设定一定阈值以清晰展示主题之间的演化路径，并通过 Graphviz 实现可视化。给定任意窗口主题 $t_h$，$t_h$ 的向量表示为 $\frac{1}{|t|}\sum_{i\in t} wv_i$，$t$ 为主题的前 $t$ 个词表示，$wv_i$ 为第 $i$ 个词的词向量。如式(4)所示，给定任意两个相邻时间窗口的窗口主题 $t_h$ 和 $t_h'$，主题之间相似度的计算可以用余弦相似度表示。对应到图形可视化上，如果两个相邻时间窗口的主题满足一定的相似度阈值，则认为这两个主题之间存在一定的演化关系，相似度越高，则这种关系就越强，在图形上的表现就是二者之间的连线越粗。

$$\mathrm{sim}(t_h, t_h') = \cos(t_h, t_h') \qquad (4)$$

本文在分析技术主题演化的时候，为清晰地展示技术主题的演化路径，分别将相邻时间窗口的两个主题之间的相似度阈值设为 0.3、0.5 和 0.7，对比不同阈值下技术主题的演化路径以选择最优阈值。如图 6 所示，当阈值设定过高时，各个时间窗口的技术主题之间的关系容易被忽略，技术主题演化轨迹不明显；当阈值设定过低时，容易引入不必要的联系，造成技术主题演化过于复杂。因此，本文将相似度阈值设定为 0.5，即如果相邻时间窗口的两个主题之间相似度大于 0.5，则认为这两个主题之间存在一定联系，相似度越高，联系越紧密，反映在演化轨迹中表现为二者之间的连线越粗。“电气设备相关技术”和“汽车生产与制造相关技术”的演化轨迹如图 7 所示。

图 7 展示了“电气设备相关技术”和“汽车生产与制造相关技术”在各个时间窗口的演化路径，每个窗口主题都用一个圆圈表示，同一列的窗口主题属于同一个时间窗口，从左至右五个时间窗口分别为 2002 年、2005 年、2008 年、2011 年和 2014 年，图中圆圈的大小通过与该主题关联的文档数量来计算，圆圈越大表明该技术主题越热门、越重要。如图 7 所示，“汽车生产与制造相关技术”共映射到 21 个窗口主题中，5 个时间窗口分别包含窗口主题 5 个、5 个、4 个、3 个和 4 个，“电气设备相关技术”共映射到 30 个窗口主题中，五个时间窗口分别包含窗口主题 5 个、6 个、3 个、7 个和 9 个。本文根据主题模型和 TextRank 名词短语抽取结果，并结合相关专利的摘要记录，提取与该动态主题相关的窗口主题释义如表 3 所示。

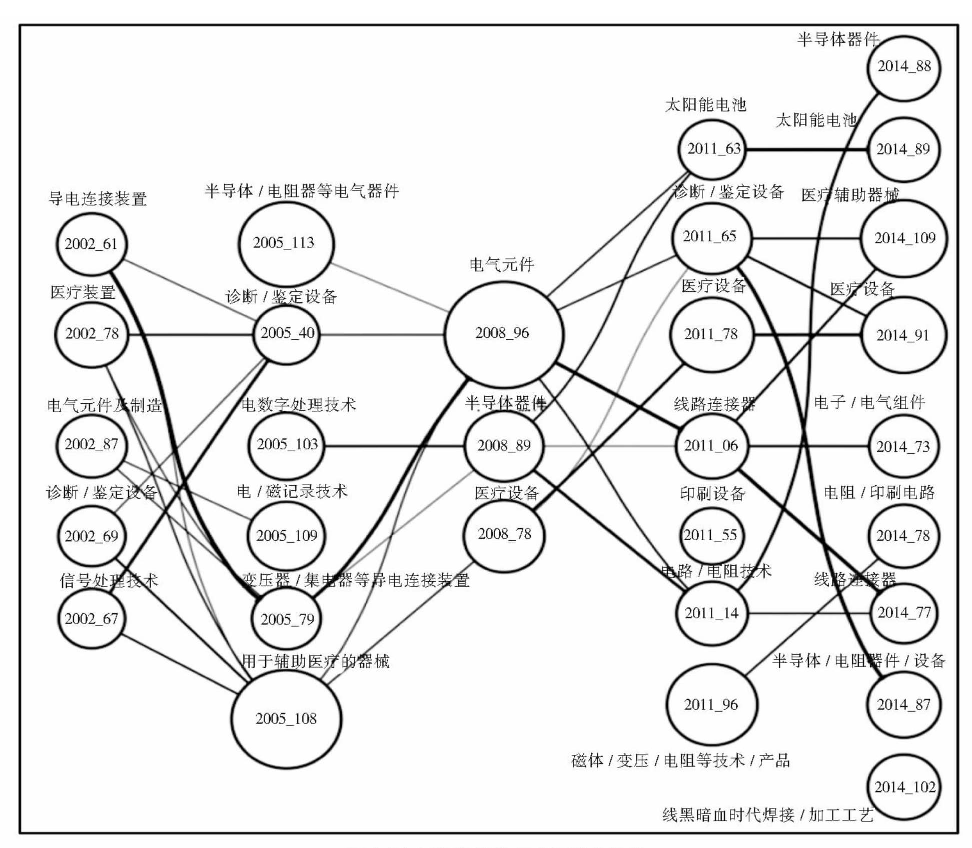
本文将窗口主题的演化模式分为新生、扩展、融合、连续和衰退五种模式[31]。新生模式表示当前主题的前一时间窗口不存在与当前主题有演化关系的主题，如图 7(1)“电气设备相关技术”演化轨迹中的“半导体/电阻器等电气器件”，“电数字处理技术”等，以及图 7(2)所示的“织物、纤维、涂料等材料技术”、2008 年的“橡胶/轮胎技术”、2014 年的“橡胶/轮胎技术”三个窗口主题。扩展模式表示当前主题与下一时间窗口的多个主题之间存在演化关系，图 7 中存在多个主题具有扩展模式，比如“汽车生产与制造相关技术”演化轨迹中 2011 年窗口主题“发动机/动力驱动技术”扩展为 2014 年“电动汽车动力及控制技术”和“发电/涡轮机技术”两个窗口主题，表明这几个技术之间有着密切地联系，“电气设备相关技术”演化轨迹中 2008 年的“电气元件”技术扩展为 2011 年的“太阳能电池”技术、“线路连接器”技术等。融合模式表示当前主题由前一时间窗口多个主题演化而来，图 7 中存在多个主题具有融合模式，比如在“汽车生产与制造相关技术”



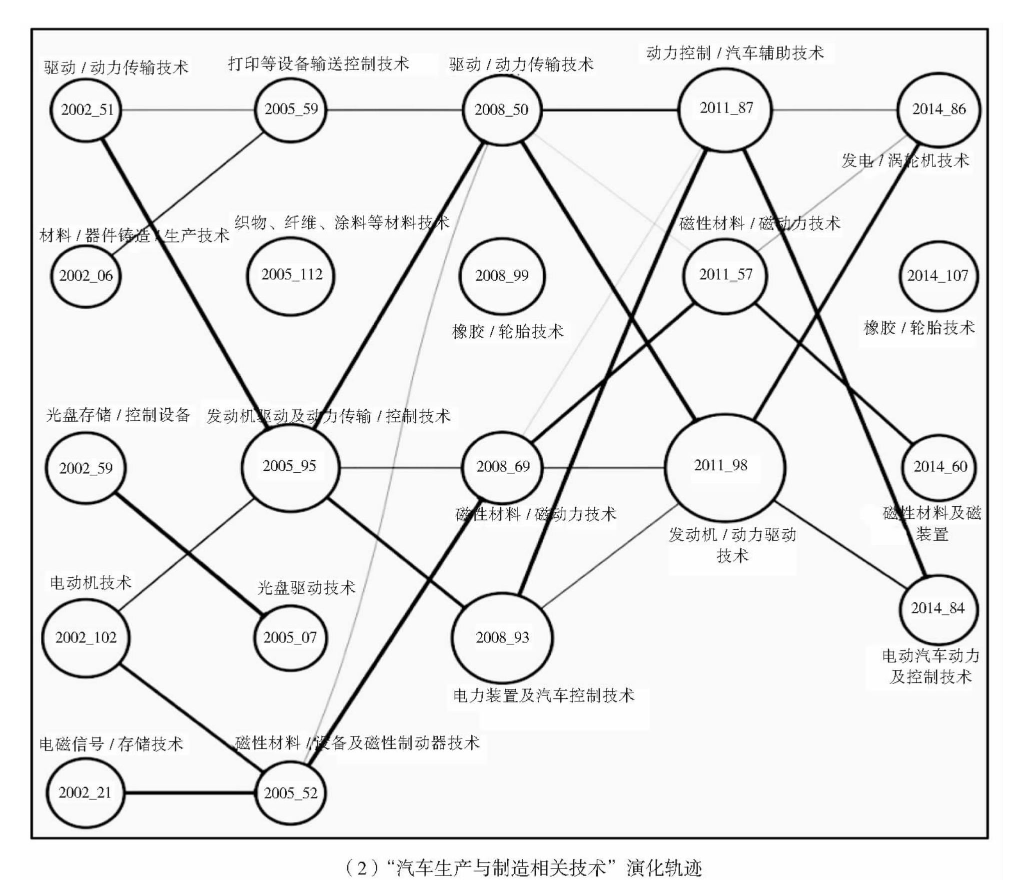
（1）“电气设备相关技术”演化轨迹，阈值从左到右分别为 0.3、0.5、0.7

（2）“汽车生产与制造相关技术”演化轨迹，阈值从左到右分别为 0.3、0.5、0.7

图 6　不同阈值的技术主题演化轨迹




（1）“电气设备相关技术”演化轨迹




（2）"汽车生产与制造相关技术" 演化轨迹

图 7　技术主题演化轨迹

表 3　技术主题释义

（1）与"电气设备相关技术"相关联的窗口主题释义

| 窗口主题 | 释义 | 窗口主题 | 释义 |
|---|---|---|---|
| 2002_61 | 导电连接装置 | 2011_65 | 诊断/鉴定设备 |
| 2002_78 | 医疗装置 | 2011_78 | 医疗设备 |
| 2002_87 | 电气元件及制造 | 2011_06 | 线路连接器 |
| 2002_69 | 诊断/鉴定设备 | 2011_55 | 印刷设备 |
| 2002_67 | 信号处理技术 | 2011_14 | 电路/电阻技术 |
| 2005_113 | 半导体/电阻器等电气器件 | 2011_96 | 磁体/变压/电阻等技术/产品 |
| 2005_40 | 诊断/鉴定设备 | 2014_88 | 半导体器件 |
| 2005_103 | 电数字处理技术 | 2014_89 | 太阳能电池 |
| 2005_109 | 电/磁记录技术 | 2014_109 | 医疗辅助器械 |
| 2005_79 | 变压器/集电器等导电连接装置 | 2014_91 | 医疗设备 |
| 2005_108 | 医疗辅助器械 | 2014_73 | 电子/电气组件 |
| 2008_96 | 电气元件 | 2014_78 | 电阻/印刷电路 |
| 2008_89 | 半导体器件 | 2014_77 | 线路连接器 |
| 2008_78 | 医疗设备 | 2014_87 | 半导体/电阻器件/设备 |
| 2011_63 | 太阳能电池 | 2014_102 | 线路焊接/加工工艺 |

（2）与"汽车生产与制造相关技术"相关联的窗口主题释义

| 窗口主题 | 释义 | 窗口主题 | 释义 |
|---|---|---|---|
| 2002_51 | 驱动/动力传输技术 | 2008_99 | 橡胶技术 |
| 2002_06 | 材料/器件铸造/生产技术 | 2008_50 | 驱动/动力传输技术 |
| 2002_59 | 光盘存储/控制设备技术 | 2008_93 | 电力装置及汽车控制技术 |

（续表 3）

| 窗口主题 | 释义 | 窗口主题 | 释义 |
|---|---|---|---|
| 2002_102 | 电动机技术 | 2011_98 | 发动机/动力驱动技术 |
| 2002_21 | 电磁信号/存储技术 | 2011_87 | 动力控制/汽车辅助技术 |
| 2005_52 | 磁性材料/设备及磁性致动器技术 | 2011_57 | 磁性材料/磁动力技术 |
| 2005_07 | 光盘驱动技术 | 2014_84 | 电动汽车动力及控制技术 |
| 2005_59 | 打印等设备输送控制技术 | 2014_107 | 橡胶/轮胎技术 |
| 2005_112 | 织物、纤维、涂料等材料技术 | 2014_60 | 磁性材料及磁装置 |
| 2005_95 | 发动机驱动及动力传输、控制技术 | 2014_86 | 发电/涡轮机技术 |
| 2008_69 | 磁材料/动力技术 | | |

的演化轨迹中，2014 年窗口主题“电动汽车动力及控制技术”主要由 2011 窗口主题“发动机/动力驱动技术”和“动力控制/汽车辅助技术”两个窗口主题融合演化而来，在“电气设备相关技术”演化轨迹中，2011 年的“太阳能电池”技术是由 2008 年的“电气元件”技术和“半导体器件”技术两个窗口主题融合演化而来。连续模式表示下一时间窗口的主体中存在且仅存在一个主题与当前主题存在演化关系，如图 7（1）所示的 2002 年窗口主题“光盘存储/控制设备技术”与 2005 年窗口主题“光盘驱动技术”之间就是连续模式，图 7（2）所示的 2011 年窗口主题“太阳能电池技术”和 2014 年窗口主题“太阳能电池技术”也属于连续模式。衰退模式表示当前主题与下一时间窗口的所有主题都不存在演化关系，如图 7（1）所示的窗口主题“光盘驱动技术”、“织物、纤维、涂料等材料技术”和“橡胶/论坛技术”。

根据图 7 所示的演化轨迹，可以分别提取出“电气设备相关技术”和“汽车生产与制造相关技术”的核心演化路径，如“电气设备相关技术”中“半导体器件”、“太阳能电池”、“医疗设备”等核心演化路径，“汽车生产与制造相关技术”中“电动汽车动力及控制相关技术”、“磁性材料及磁装置相关技术”、“发电/涡轮机相关技术”等核心演化路径。以“汽车生产与制造相关技术”中三个核心演化路径为例：①“电动汽车动力及控制相关技术”，演化路径如图 8（1）所示，“电磁技术”和“电动机技术”通过融合模式演化为“电磁致动器技术”，“电动机技术”与“驱动/动力传输技术”通过融合模式演化为“发动机驱动及动力传输、控制技术”，到了 2008 年，“发动机驱动及动力传输、控制技术”又通过扩展模式演化为“动力驱动/传输技术”和“电力装置及汽车控制技术”，“磁性材料/设备及磁性致动器技术”则通过连续模式演化为“磁材料/动力技术”，之后，“磁材料/动力技术”和“动力驱动/传输技术”以及“动力驱动/传输技术”和“电力装置及汽车控制技术”分别通过融合模式演化为“发动机/动力驱动技术”以及“动力控制/汽车辅助技术”，最后这两个技术通过融合模式演化为“电动汽车动力及控制相关技术”；②“磁性材料及磁装置相关技术”演化路径如图 8（2）所示，“电磁技术”和“电动机技术”通过融合模式演化为“电磁致动器技术”，然后一直通过连续模式演化为“磁性材料及磁装置技术”；③“发电/涡轮机相关技术”，演化路径如图 8（3）所示，“电磁技术”和“电动机技术”通过融合模式演化为“电磁致动器技术”，之后通过连续模式演化为“磁材料/动力技术”，“驱动/动力传输技术”通过连续模式演化为“动力驱动/传输技术”，之后与“磁材料/动力技术”通过融合模式演化为“发动机/动力驱动技术”，最后通过连续模式演化为“发电/涡轮机技术”。虽然本文将这三个核心技术的演化轨迹分别抽取出来分析，但三者的演化轨迹存在着很多交叉重叠的现象，表明这三个核心技术之间仍然存在着紧密的联系。在这三个核心技术演化轨迹之外的其他技术则较为孤立，如 2014_107 和 2008_99 表示的橡胶技术（汽车轮胎的橡胶材料），2005_112 表示的织物、涂料等功能材料技术等。这些技术在动态主题 D64 所表示的汽车动力与制造相关技术中处于非核心的位置，而且很少与其他技术产生联系，演化轨迹不明显。

## 5 总结与展望

本文从文本语义的角度出发，提出基于 NMF 改进的动态非负矩阵分解模型，对专利文本进行动态主题建模，以实现对技术主题的动态演变分析，主要分为五个步骤：①通过 Word2Vec 训练词向量获取主题词的分布表示，用于主题模型中主题个数 k 的确定以及主题之间相似度的计算；②通过改进的动态非负矩阵分解对专利文本进行动态主题建模，获取动态主题及相对

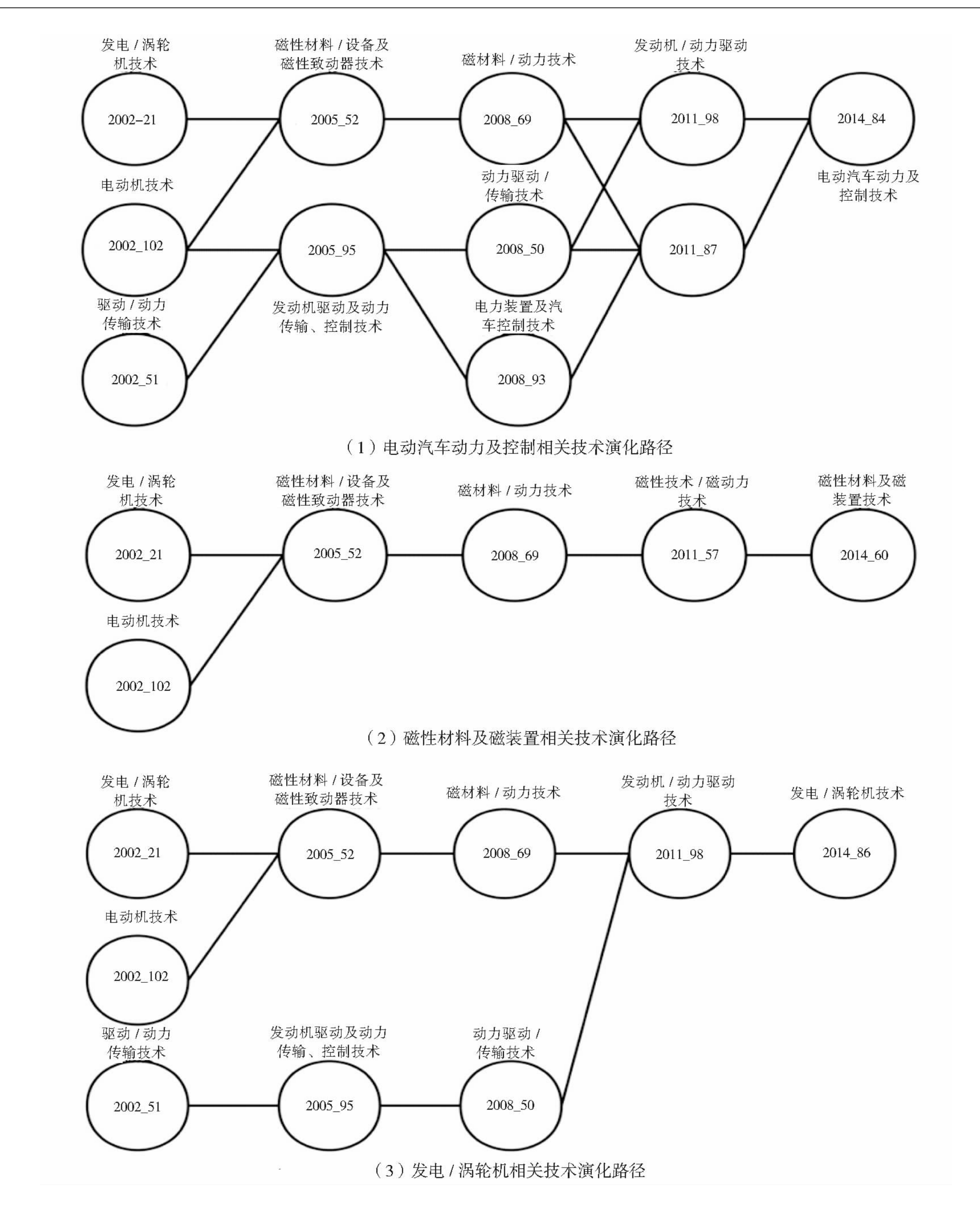



图 8　三个核心技术的演化路径

应的窗口主题;③利用 TextRank 抽取名词短语对抽取的主题进行标注,增强主题的可解释性;④通过词向量计算主题之间的演化轨迹,并通过 Graphviz 可视化展示;⑤选取 2002 年、2005 年、2008 年、2011 年和 2014 年五年的五方专利数据进行实证分析。

本文提出的方法能够充分利用专利的文本内容信息,自动识别专利文本中蕴含的技术主题,并识别其演化路径。本文仍存在一些不足之处亟待进一步研究:①本文通过先抽取主题然后再利用名词短语进行标注的方式增强主题的可解释性,虽然达到了一定的效果,但处理步骤较为繁琐,研究如何直接生成主题短语的主题模型成为笔者未来的研究重点;②本文在进行主题建模的时候,只用到了文本信息,而忽略了专利分类等有价值的信息,因此如何将专利分类等信息融入主题模型以提升主题模型的精度和效果也需要进一步研究。

**参考文献:**

[1] ROSENBERG N. Technological change in the machine tool industry, 1840－1910 [J]. The journal of economic history, 1963, 23 (4): 414－443.

[2] ROSENBERG N. Exploring the black box: technology, economics, and history [M]. Cambridge: Cambridge University Press,

1994.

[3] CHO Y, KIM M. Entropy and gravity concepts as new methodological indexes to investigate technological convergence: patent network-based approach [J/OL]. PLOS ONE, 2014, 9(6): e98009 [2016-10-26]. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098009.

[4] 胡阿沛，张静，雷孝平，等. 基于文本挖掘的专利技术主题分析研究综述[J]. 情报杂志, 2013(12): 88-92.

[5] 肖沪卫. 专利地图方法与应用[M]. 上海:上海交通大学出版社, 2011.

[6] SUZUKI K, SAKATA J, HOSOYA J. An empirical analysis on progress of technology fusion [C] //IEEE. Third International conference on digital information management (ICDIM). Piscataway: IEEE Xplore, 2008: 937-939.

[7] JEONG S, KIM J C, CHOI J Y. Technology convergence: what developmental stage are we in? [J]. Scientometrics, 2015, 104(3): 841-871.

[8] LEE W S, HAN E J, SOHN S Y. Predicting the pattern of technology convergence using big-data technology on large-scale triadic patents [J]. Technological forecasting and social change, 2015, 100:317-329.

[9] HUANG B, HUANG L. Patent-based association analysis for technological convergence of IT and BT [C]//IEEE. 2015 International Conference on Logistics, Informatics and Service Sciences (LISS). Piscataway: IEEE Xplore, 2015: 1-6.

[10] NARIN F. Patent bibliometrics [J]. Scientometrics, 1994, 30(1): 147-155.

[11] CHANG P L, WU C C, LEU H J. Using patent analyses to monitor the technological trends in an emerging field of technology: a case of carbon nanotube field emission display [J]. Scientometrics, 2010, 82(1): 5-19.

[12] CHOI C, PARK Y. Monitoring the organic structure of technology based on the patent development paths [J]. Technological forecasting and social change, 2009, 76(6): 754-768.

[13] GEUM Y, KIM C, LEE S, et al. Technological convergence of IT and BT: evidence from patent analysis [J]. ETRI journal, 2012, 34(3): 439-449.

[14] 翟东升，蔡力伟，张杰，等. 基于专利的技术融合创新轨道识别模型研究——以云计算技术为例[J]. 情报学报, 2015, 34(4): 352-360.

[15] MOGEE M E, KOLAR R G. Patent co-citation analysis of Eli Lilly & Co. patents [J]. Expert opinion on therapeutic patents, 1999, 9(3): 291-305.

[16] HUANG M H, CHIANG L Y, CHEN D Z. Constructing a patent citation map using bibliographic coupling: a study of Taiwan's high-tech companies [J]. Scientometrics, 2003, 58(3): 489-506.

[17] 陈亮，张志强. 技术演化研究方法进展分析[J]. 图书情报工作, 2012, 56(17): 59-66.

[18] MARTINELLI A. An emerging paradigm or just another trajectory? Understanding the nature of technological changes using engineering heuristics in the telecommunications switching industry [J]. Research policy, 2012, 41(2): 414-429.

[19] MINA A, RAMLOGAN R, TAMPUBOLON G, et al. Mapping evolutionary trajectories: applications to the growth and transformation of medical knowledge [J]. Research policy, 2007, 36(5): 789-806.

[20] SU H N, LEE P C. Dynamic and quantitative exploration on technology evolution mechanism: the case of electrical conducting polymer nanocomposite [C]//IEEE. 2009 Portland international conference on management of engineering & technology (PICMET). Piscataway: IEEE Xplore, 2009: 2433-2440.

[21] BATAGELJ V. Analyzing the structure of US patents network [M]//Data science and classification. Springer, Berlin, Heidelberg, 2006: 141-148.

[22] ABBAS A, ZHANG L, KHAN S U. A literature review on the state-of-the-art in patent analysis [J]. World patent information, 2014, 37(4): 3-13.

[23] 栾春娟. 基于专利共现的全球太阳能技术网络及关键技术演进分析[J]. 情报学报, 2013, 32(1): 68-79.

[24] 韩红旗，安小米，朱东华，等. 专利技术术语共现的战略图分析方法[J]. 计算机应用研究, 2011, 28(2): 576-579.

[25] CHEN S H, HUANG M H, CHEN D Z. Identifying and visualizing technology evolution: a case study of smart grid technology [J]. Technological forecasting & social change, 2012, 79(6): 1099-1110.

[26] LEE D D, SEUNG H S. Learning the parts of objects by non-negative matrix factorization [J]. Nature, 1999, 401: 788-791.

[27] WANG Q, CAO Z, XU J, et al. Group matrix factorization for scalable topic modeling [C]// Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2012: 375-384.

[28] BOUTSIDIS C, GALLOPOULOS E. SVD based initialization: a head start for nonnegative matrix factorization [J]. Pattern recognition, 2008, 41(4): 1350-1362.

[29] CHANG J, GERRISH S, WANG C, et al. Reading tea leaves: how humans interpret topic models [C]// Proceedings of the 22nd international conference on neural information processing systems. USA: Curran Associates Inc., 2009: 288-296.

[30] O'CALLAGHAN D, GREENE D, CARTHY J, et al. An analysis of the coherence of descriptors in topic modeling [J]. Expert systems with applications, 2015, 42(13): 5645-5657.

[31] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Proceedings of the 26th international conference on neural information processing systems. USA: Curran Associates Inc., 2013: 3111-3119.

[32] 潘教峰，张晓林，王小梅，等. 科学结构地图 2012 [M]北京:

科学出版社,2013.

[33] SULO R, BERGER-WOLF T, GROSSMAN R. Meaningful selection of temporal resolution for dynamic networks [C]//Proceedings of the Eighth workshop on mining and learning with graphs. New York: ACM, 2010: 127 - 136.

[34] LEE H J, LEE S, YOON B. Technology clustering based on evolutionary patterns: the case of information and communications technologies [J]. Technological forecasting and social change, 2011, 78(6): 953 - 967.

[35] KIM E, CHO Y, KIM W. Dynamic patterns of technological convergence in printed electronics technologies: patent citation network [J]. Scientometrics, 2014, 98(2): 975 - 998.

[36] 李姝影, 方曙. 测度技术融合与趋势的数据分析方法研究进展[J]. 数据分析与知识发现, 2017, 1(7): 2 - 12.

[37] HAN E J, SOHN S Y. Technological convergence in standards for information and communication technologies [J]. Technological forecasting and social change, 2016, 106: 1 - 10.

[38] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation [J]. Journal of machine learning research, 2003, 3: 993 - 1022.

[39] HOFMANN T. Unsupervised learning by probabilistic latent semantic analysis [J]. Machine learning, 2001, 42(1 - 2): 177 - 196.

[40] MEI Q, SHEN X, ZHAI C. Automatic labeling of multinomial topic models [C]//Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2007: 490 - 499.

[41] EL-KISHKY A, SONG Y, WANG C, et al. Scalable topical phrase mining from text corpora [J]. Proceedings of the VLDB endowment, 2014, 8(3): 305 - 316.

[42] WANG X, MCCALLUM A, WEI X. Topical n-grams: phrase and topic discovery, with an application to information retrieval [C]//Proceedings of the 2007 Seventh IEEE international conference on data mining. Washington, DC :IEEE Computer Society, 2007: 697 - 702.

[43] BLEI D M, LAFFERTY J D. Visualizing topics with multi-word expressions [EB/OL]. [2016 - 10 - 28] https://arxiv.org/abs/0907.1013.

[44] MIHALCEA R, TARAU P. TextRank: bringing order into texts [C]// Proceedings of the ACL 2004 on interactive poster and demonstration sessions. Stroudsburg: Association for Computational Linguistics, 2004: 20.

**作者贡献说明:**

王园园:负责方法的提出,实验的设计与论文撰写;

赵亚娟:负责整体方向的把握和论文的定稿。



## Evolution Analysis of Technological Topic: An Approach Based on Non-negative Matrix Factorization

Wang Yuanyuan　Zhao Yajuan

National Science Library, Chinese Academy of Sciences, Beijing 100190

School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

**Abstract:** [**Purpose/significance**] Analyzing the evolution of technological topic makes it possible for us to track the development of technology, which is essential for improving innovation activity and forecasting development trends of technology. However, to our knowledge, scholars pay less attention to the semantic perspective of technological topic. Therefore, this paper intends to analyze the evolution of technological topic from the perspective of semantic. [**Method/process**] This paper proposed a dynamic topic model based on non-negative matrix factorization, and labeled the technology topics with noun phrases extracted by TextRank algorithm, which enhances the interpretability. Then, the study computed and visualized the evolutionary trajectory of technological topics with word embedding. [**Result/conclusion**] This paper uses five countries' (China, America, Japan, South Korea, Europe) patent data in 2002, 2005, 2008, 2011 and 2014 to test our model. During the course of the experiment, our method extracted evolutionary trajectories of 65 technical topics, which verified the effectiveness of our method.

**Keywords:** technological topic evolution　non-negative matrix factorization(NMF)　topic model　dynamic topic analysis